SELF-RAG Evaluation Project

Ayala Socolovsky, Bar Gazit, Fouzi Abdallah, Stas Rodov

July 7th, 2024

Ben-Gurion University of the Negev

Table of Contents

Original Paper: SELF-RAG

SELF-RAG: LEARNING TO RETRIEVE, GENERATE, AND CRITIQUE THROUGH SELF-REFLECTION

Akari Asai[†], Zeqiu Wu[†], Yizhong Wang[†§], Avirup Sil[‡], Hannaneh Hajishirzi[†§]
[†]University of Washington [§]Allen Institute for AI [‡]IBM Research AI
{akari,zeqiuwu1,yizhongw,hannaneh}@cs.washington.edu, avi@us.ibm.com

Reflection Tokens

Type	Input	Output	Definitions
Retrieval	x / x, y	{yes, no, continue}	Decides when to retrieve with $\mathcal{R}$
IsRel	x, d	{**relevant**, irrelevant}	d provides useful information to solve x.
IsSup	x, d, y	{**fully supported**, partially supported, no support}	All of the verification-worthy statement in y is supported by d.
IsUse	x, y	{**5**, 4, 3, 2, 1}	y is a useful response to x.

Four types of reflection tokens used in SELF-RAG. Each type uses several tokens to represent its output values. The bottom three rows are three types of Critique tokens, and **the bold text** indicates the most desirable critique tokens. x, y, d indicate input, output, and a relevant passage, respectively.

Algorithm

Algorithm 1 SELF-RAG Inference

Require: Generator LM $\mathcal{M}$, Retriever $\mathcal{R}$, Large-scale passage collections $\{d_1, \ldots, d_N\}$

Input: input prompt x and preceding generation $y_{<t}$, **Output:** next output segment y_t

$\mathcal{M}$ predicts $\boxed{\text{Retrieve}}$ given $(x, y_{<t})$

if $\boxed{\text{Retrieve}}$ == `Yes` **then**

  Retrieve relevant text passages **D** using $\mathcal{R}$ given (x, y_{t-1}) ▷ Retrieve

  $\mathcal{M}$ predicts $\boxed{\text{IsRel}}$ given x, d and y_t given $x, d, y_{<t}$ for each $d \in$ **D** ▷ Generate

  $\mathcal{M}$ predicts $\boxed{\text{IsSup}}$ and $\boxed{\text{IsUse}}$ given x, y_t, d for each $d \in$ **D** ▷ Critique

  Rank y_t based on $\boxed{\text{IsRel}}$, $\boxed{\text{IsSup}}$, $\boxed{\text{IsUse}}$ ▷ Detailed in Section 3.3

else if $\boxed{\text{Retrieve}}$ == `No` **then**

  $\mathcal{M}_{gen}$ predicts y_t given x ▷ Generate

  $\mathcal{M}_{gen}$ predicts $\boxed{\text{IsUse}}$ given x, y_t ▷ Critique

RAG vs Self-RAG

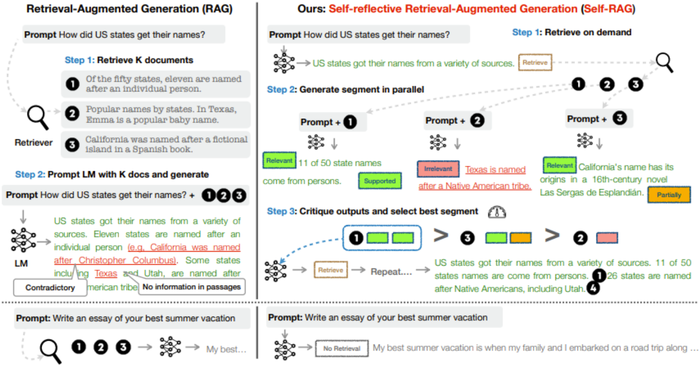

Figure 1: Overview of SELF-RAG. SELF-RAG learns to retrieve, critique, and generate text passages to enhance overall generation quality, factuality, and verifiability.

Project Objective

Evaluate $\boxed{\text{IsRel}}$ and $\boxed{\text{Retrieval}}$ tags quality

Table of Contents

Dataset: Attributed QA

- Based on Google's Natural Questions (NQ) dataset
- Contains trivia questions written by real Google users
- **Requires knowledge** (e.g., information found on Wikipedia)

Columns:

- Question
- Passage retrieved by a retriever for that question
- Human annotation indicating whether the passage is relevant to the question
- Other, less relevant, columns like "Generated answer" and etc.

Setup

- For the entire project we used Google Colab to load Self-RAG system from Huggingface
- Google Colabs L4 and A100 machines were used for 7B and 13B models respectively

Evaluating “IsRelevant” Tags Family

Called Self-RAG iterativly with the data from Attributed QA dataset:

- Cleaned data - removed duplicate and contradictions
- Question + its respective passage
- Save Self-RAG response + check whether it contains [Relevant] or [Irrelevant] tags
- Compare the result to the human annotation from the dataset (indicating whether the passage is relevant to the question or not)

Imitating “IsRelevant” Tags Family - Claude

Comparison with state-of-the-art model:

- Chose Claude 3.5 Sonnet for comparison
- One of the most powerful models globally
- Competes and sometimes surpasses GPT-4 in various tasks

Imitating "IsRelevant" Tags Family - Claude

Claude 3.5 Sonnet evaluation process:

- Used Google Colab notebook for API calls
- Provided the model with the question and the passage
- Asked to determine passage relevance to the question
- Compared results with human annotation from the dataset

Claude 3.5 Sonnet Prompt

```
USER
Decide if the following passage <Passage> is relevant for answering the following question
<Question>.
Explain your reasoning process, then write the final answer "Yes" or "No" between the tags
<Answer> and </Answer>.
Here is the question <Question>: <Question>when does like cage season 2 come out</Question>
Here is the passage <Passage>: <Passage>Title: Luke Cage (season 2)
Section: Release

The second season of Luke Cage was released on June 22, 2018, on the streaming service
Netflix worldwide, in Ultra HD 4K and high dynamic range.</Passage>
```

Imitating “IsRelevant” Tags Family - Gemma 2 27B

- Free and open SOTA model
- Used RunPod machine with A100 PCIe 80GB
- Evaluation process same as with Claude

Gemma 2 27B Prompt Template

```
<Q>{question_1}</Q> <P>{passage_1}</P> <A>{human_rating_YES}</A>
<Q>{question_2}</Q> <P>{passage_2}</P> <A>{human_rating_YES}</A>
<Q>{question_3}</Q> <P>{passage_3}</P> <A>{human_rating_NO}</A>
<Q>{question_4}</Q> <P>{passage_4}</P> <A>{human_rating_NO}</A>
<Q>{question_INPUT}</Q> <P>{passage_INPUT}</P> <A>
```

Evaluating "Retrieval" Tags Family

Decided to continue to another experiment - This time on **Retrieval** family

Working hypothesis:

- Questions from Attributed QA dataset are trivia questions
- We expect a model trained to "choose" whether external information is needed, to generate "Retrieve" token for most, if not all, questions of this type

Evaluating “Retrieval” Tags Family

Called Self-RAG iterativly with questions from Attributed QA dataset:

- Save Self-RAG response + check whether it contains [Retrieval] or [No Retrieve] tags
- Compare the result to our expectations of high percentage of [Retrieval] and low percentage of [No Retrieve]

Imitating "Retrieval" Tags Family - Claude

Used Claude 3.5 Sonnet again:

- Provided the model with the question
- Asked to determine if, given a question, there's a need to retrieve information from Wikipedia
- Compared results with our expectations

Claude 3.5 Sonnet Prompt

SYSTEM PROMPT

Pretend that you have access to all Wikipedia, so you can use Wikipedia to answer trivia questions

USER

Decide if you would use Wikipedia to answer the following question <Question>.
Write the final answer "Yes" or "No" between the tags <Answer> and </Answer>.
Here is the question <Question>: <Question>who played hyde in league of extraordinary gentlemen</Question>

Table of Contents

Results - “IsRelevant” Tags Family

49.9% match with human rating

		Human Rating	
		Relevant	Irrelevant
Self-RAG 7B	Relevant	2087 (36.1%)	2869 (49.6%)
	Irrelevant	31 (0.5%)	798 (13.8%)

46.8% match with human rating

		Human Rating	
		Relevant	Irrelevant
Self-RAG 13B	Relevant	2080 (36.0%)	3042 (52.6%)
	Irrelevant	38 (0.7%)	625 (10.8%)

Confusion Matrices: Self-RAG 7B, 13B

Results - Imitating "IsRelevant" Tags Family

76.4% match with human rating

		Human Rating	
		Relevant	Irrelevant
Claude 3.5 Sonnet*	Relevant	1932 (33.4%)	1181 (20.4%)
	Irrelevant	186 (3.2%)	2486 (43.0%)

65.34% match with human rating

		Human Rating	
		Relevant	Irrelevant
Gemma 2**	Relevant	467 (16.87%)	279 (10.08%)
	Irrelevant	676 (24.41%)	1342 (48.47%)

Confusion Matrices: Gemma-2 and Claude vs Human Ratings

* Checked 100 examples **without** CoT- results were almost identical

** Gemma-2 might have seen the whole dataset during training

Results - "Retrieval" Tags Family

	Retrieval	No Retrieval	No tags
Self-RAG 13B	341 (5.9%)	36 (0.6%)	5408 (93.5%)
Claude 3.5 Sonnet	93 (93%)	7 (7%)	-

Table: Retrieval Decision Comparison: Self-RAG vs Claude

Table of Contents

"IsRelevant" Results Analysis

- Self-RAG models have strong bias towards generating "Relevant" tag, at least for Attributed QA dataset

Possible explanations for the **very** poor results:

- User may need to use the same retriever used in the fine-tuning step to avoid poor results
- Maybe more documents per question might increase chances of a good result getting good score
- Problem definition: the task of determining whether a document is relevant to answer a question may not be an easy one, for humans as well.

"IsRelevant" Results Analysis

Future research:

- Check performance of Llama 2 model (Self-RAG base model) on Attributed QA dataset
- Maybe Self-RAG fine-tuning actually reduced Llama 2 ability to determine relevance

"Retrieval" Results Analysis

- The model did not behave as expected
- Absence of a $\boxed{\textbf{Retrieval}}$ tag may be interpreted as "No retrieval" (wasn't mentioned in the paper)
- However, even in this scenario, it still contrary to our expectations
- Maybe the threshold for "Retrieval" tags should be lower

Table of Contents

Conclusion

- Don't believe to everything you read!
- Claude 3.5 Sonnet is very powerful language model that competes GPT-4
- "Here's the English translation of your Hebrew text, formatted as LaTeX bullet points"

Code

https://github.com/stasrodov/self-rag-eval